DOCUMENT RESUME

ED 430 006 TM 029 719

AUTHOR Mott, Michael S.; Halpin, Regina

TITLE A Comparative Approach to the Examination of the Technical Qualities of a Pen and Paper Writing Assessment for Elementary Students' Hypermedia-Created Products.

PUB DATE 1999-04-00

NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999). For the authored narratives, see TM 029 720.

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Comparative Analysis; *Computer Assisted Instruction; Computer Software; Elementary School Students; Evaluation Methods; *Hypermedia; Primary Education; Reliability; Test Construction; Test Format; *Test Use; Validity

IDENTIFIERS Paper and Pencil Tests; Scoring Rubrics; *Writing What You Read

ABSTRACT

The reliability and developmental and concurrent validity of the Writing What You Read (WWYR) rubric, designed for use with paper and pen, for hypermedia-authored narrative productions of students in grades 2 and 3 were studied. Sixty students from 4 classrooms produced hypermedia narratives (interactive multimedia presentations) that were rated by 5 trained raters (doctoral students) with knowledge in the teaching of writing process and the use of hypermedia software. The study used an ex post facto design with a comparative component to examine the reliability and developmental validity of the WWYR rubric through percentages of agreements and Pearson correlations. Student achievement was also assessed through the Iowa Tests of Basic Skills (ITBS), and scores from the ITBS and the WWYR were positively correlated. Other reliability and validity analyses suggest that teachers may benefit from applying the WWYR assessment to their students' hypermedia productions. (Contains 7 tables, 3 figures, and 19 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Running head: COMPARATIVE APPROACH

ED 430 006

A Comparative Approach to the Examination of the Technical Qualities of a Pen and Paper Writing Assessment for Elementary Students' Hypermedia-Created Products

Michael S. Mott

National-Louis University

Regina Halpin

Mississippi State University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Michael Mott

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Paper Presented at the 1999 Annual Meeting of the

American Educational Research Association

Montreal, Canada

April 19-23

TM029719

BEST COPY AVAILABLE

Abstract

This study[1] investigated the reliability, and the developmental and concurrent validity, of the Writing What You Read (WWYR) rubric for hypermedia-authored narrative productions of students in grades 2 and 3. Students (n=60) from four intact classrooms produced hypermedia narratives (interactive multimedia presentations that consisted of text, graphics and audio elements) over four months in a school-based computer laboratory equipped with ten Windows-based microcomputers. Raters (n=5) with knowledge in the teaching of process writing and use of hypermedia software judged the hypermedia narrative productions. The researcher developed an interactive hypermedia software tutorial program which was used to train teachers (n=4) how to implement a process writing/hypermedia curriculum.

Raters participated in a three-hour training and rating session in a university computer laboratory equipped with five Power Macintosh microcomputers. Raters judged all students' (n=60) hypermedia narrative productions individually without resolving differences through discussion.

This study used an ex post facto design with a comparative component to examine the reliability and developmental validity of the WWYR rubric for scoring hypermedia-created narrative productions. Two analyses were used to determine reliability: percentages of agreement and Pearson correlations. Percentages of agreement for the WWYR Rubric averaged across ten pairs of raters found high percentages of agreement among raters (.70 for ±0 and .99 for ±1). Pearson correlations averaged across ten pairs of raters found acceptable interrater reliability for four (Theme, Character, Plot and Communication) of the five subscales. (For Theme, Character, Setting, Plot and Communication the r values were .59, .55, .49, .50 and .50).

Developmental validity of the WWYR scores were examined in two ways. First, Hotelling's T^2 was used to compare the ratings assigned to productions of students in grade 2 with the ratings of students in grades 3. No significant differences were observed. Second, One-Way MANOVA was used to evaluate WWYR scores of students grouped as low, medium or high ability based on their Iowa Test of Basic Skills (ITBS) National Percentile Rank for Literacy skill. A statistically significant difference was observed between mean vectors across the three ability groups, $F(2, 36)=2.59$, $p<.01$. Concurrent validity was examined through correlational analysis, between students' mean WWYR score and ITBS score. Scores from the two measures were positively correlated, $r=.83$, $p<.01$. Results, from both, the One-Way MANOVA and Correlational analysis provided evidence for the score sensitivity of the WWYR assessment to the developmental literacy competency of the grade 3 students.

[1] The current technical paper is based on Mott (1998), an unpublished dissertation. Mott and Hare (1999) provides a different view of the current study by placing results in a curriculum and instruction context.

Introduction

In recent years, a considerable amount of research activity has occurred in the following three areas: (a) direct writing assessment, (b) process writing; and (c) learning and writing in hypermedia computer environments. Much of the activity concerning direct writing assessment has focused on the reliability and validity of raters' judgments of writing quality. (Figure 1 summarizes this research activity). Several studies have examined and supported the reliability and validity of raters' judgments utilizing rubrics to measure the quality of writing samples created on paper by elementary students (Gearhart, Herman, Novak, & Wolf, 1995; Novak, Herman, & Gearhart, 1996). Rubrics such as Writing What You Read (WWYR) shown in Table 1 has been examined in recent reliability and validity studies and used as assessments of student writing samples created within a process writing curriculum such as Writing Workshop. The Writing Workshop consists of students engaged in a process of writing consisting of numerous cycles of the discrete stages: brainstorming, editing, publishing, etc (Graves, 1983). While these studies have addressed the assessment of pen and paper-created writing samples, efforts to develop direct assessment rubrics for evaluating process outcomes created in hypermedia environments have largely been neglected.

This study examined the relationship between hypermedia-created narrative products and raters' judgments of quality using the WWYR, a direct assessment protocol previously evaluated for pen and paper-created writing samples. Specifically, this study attempted to establish the degree of interrater reliability, and developmental and concurrent validity, of raters' judgment scores based on the quality of students' hypermedia-created productions. Developmental validity represents the sensitivity of the WWYR assessment to detect differences in grade and ability levels (Figure 2 places developmental validity into a meaningful context). Concurrent validity represents the degree to which scores on the WWYR are related to scores on an already established test, the Iowa Test of Basic Skills (ITBS). These two validity types must be established in order for a measure to attain content-related validity.

Purposes and Hypotheses

The purpose of this study was to address an emerging concern in the field of writing assessment and hypermedia learning regarding the need for a vehicle to reliably and validly assess students' hypermedia-created productions. Researching this issue represented one step in the process of evaluating

the reliability and validity of an assessment that could eventually be used to evaluate the impact of hypermedia writing on student learning. Three null hypotheses were tested:

1. There will not be acceptable levels of interrater reliability among raters' assessment scores based on the Writing What You Read (Wolf & Gearhart, 1993a, 1993b) analytic subscales of Theme, Character, Setting, Plot and Communication when they are used to evaluate hypermedia created narrative productions of students in grades 2 and 3.

2. There will be no significant difference (=0.05) between WWYR mean vectors (based on the five analytic subscale scores of Theme, Character, Setting, Plot and Communication) for hypermedia-created narratives productions of second grade and third grade students.

3. There will be no significant difference (=0.05) between WWYR mean vectors (based on the five analytic subscale scores of Theme, Character, Setting, Plot, and Communication) for hypermedia-created narrative productions of third grade students classified as low, medium, or high ability based on their scores on the ITBS (Linn & Willson, 1990).

Theoretical Framework

According to Ayersman (1996), student-created hypermedia documents containing presentations with any combination of text, hypertext, graphics, audio and video which focused on disciplinary topics, can enhance learning since this environment supports constructivist theory. These hypermedia attributes (text, hypertext, graphics, audio and video) were identified as features conducive to the teaching of writing. Swan and Meskill (1996) found hypermedia to be a potentially suitable environment for literacy learning that included support for: (a) independent learning, (b) cooperative learning, (c) non-linear representations of knowledge, (d) a wide array of learning styles, and (e) enabling teachers to evaluate their own ideas of the role of text in the teaching of writing and reading.

McLellan (1992), in case study research of a hypermedia writing curriculum, investigated how elementary students (grade 5) would excel in narrative writing in the HyperCard environment. Students developed their own stories and manipulated the non-linear hypertextual features of the software. The level of details were strengthened in both narrative and episodic story structures, and McLellan noted that the children quickly adapted to the hypermedia environment. Smith (1992) engaged Navajo elementary boarding school children (grades 3-6) in the implementation of the hypermedia authoring software Linkway

which supported the integration of text, audio, video and graphics for IBM compatible computers, similar to Hyperstudio (Wagner, 1995) for Macintosh operating environments.

The proliferation and increasing popularity among elementary school teachers of hypermedia learning environments, particularly for writing, dictate that research needs to address this new frontier. (See Figure 3 for a summary of hypermedia writing environments). Several researchers have expressed the desire that new research for hypermedia writing products be developed (Kinzer & Leu, 1997; Palumbo & Prater, 1993; Reed, 1996; Sharp, Kinzer & Risko, 1994; Yang, 1996). Palumbo and Prater (1993) and Ayersman (1996) further related that new assessment research is especially necessary in order to facilitate the development of writing instruction that makes effective use of hypermedia. Gearhart et al. (1995) concluded, that writing assessment research is needed to determine the factors that support or constrain the judgments of popular and extensively researched writing rubrics, particularly the WWYR analytic/holistic rubric. Thus, there is a particular need for writing assessment research to be conducted on the types of material to be rated, such as hypermedia documents instead of pen and paper-created documents.

Instrumentation

The WWYR Rubric (see Table 1) contains five evaluative scales designed to assess students' developing competencies in narrative writing: Theme, Character, Setting, Plot and Communication. The vertical analytical evaluative scales (1-6 for each competency) were designed to enable teachers to make instructional decisions on specific narrative components a student needs reinforcement in, and were not intended as a method for assigning a numerical value to a narrative. Teachers merely shade off a box in the rubric to denote where a child's narrative is along each competency. The ITBS (Linn & Wilson, 1990) Form J was used as a basic battery for grades k-9 and includes language skills directly related to writing: word analysis, vocabulary, spelling and reading comprehension. Reliability coefficients for Form J ranged from .70-.90 for the language skills components. Additionally the ITBS meets high standards of overall technical quality and is a widely accepted standardized measure of cognitive skill.

Procedures

Three data sources were employed in the study: students, teachers and raters. Four teachers who were knowledgeable in process writing curricula and Hyperstudio hypermedia software received additional training in both process writing and hypermedia software use. Sixty students from grades 2 and 3 created,

with the assistance of their teachers, hypermedia narrative products in Hyperstudio (Wagner, 1995) as part of a four month-long curriculum. The five raters, all doctoral students experienced in process writing curriculum and hypermedia software applications, were trained on the WWYR narrative rubric by the researcher in a three-hour training and rating session. An ITBS National Percentile Rank for literacy competency was obtained for each student in grade 3. This data was used to examine the developmental and concurrent validity of WWYR scores. One-Way MANOVA revealed significant differences across all five WWYR subscale scores between students' classified as low, medium or high ability (ITBS).

Results for Percentages of Agreement and Pearson Correlations

Percentages of Agreement

An examination of the percentages of agreement for the WWYR rubric assessment scores averaged across ten pairs of raters and the Pearson correlations for WWYR rubric scores averaged across ten pairs of raters revealed acceptable levels of interrater reliability. Therefore, Null Hypothesis 1 was rejected.

Table 2 contains results of the percentages of agreement across all rater pairs for the five WWYR subscales. Table 3 contains results of the percentages of agreement across all rater pairs in the current study as well as the percentages of agreement observed for two other WWYR studies (Gearhart et al. 1995; Novak et al. 1996).

The results for the current study indicated that the ±0 and ±1 percentages of agreement across ten pairs of raters were higher than the ±0 and ±1 agreement levels found in both the Gearhart et al. (1995) and Novak et al. (1996) WWYR reliability studies. The high percentages of agreement found in this study may be attributed to the raters' use of only the first three WWYR rubric evaluative subscale levels. The WWYR rubric contains six subscale levels that are developmentally sequenced according to the varied writing competencies of students in grades K-6. Since students in this study were in grades 2 and 3, only levels 1, 2 and 3 were typically applied by the raters when judging the hypermedia narrative productions. This narrow range of values independently applied by raters functioned to limit the number of choices. Hence, high percentages of agreement between raters would be expected based on the limited number of scale levels used.

The percentages of agreement that were revealed in the current study, although higher than those found in the Gearhart et al. study, should be considered descriptive information. Gearhart et al. remarked that percentages of agreement found for the WWYR should not be interpreted as "strong evidence of reliability" (p. 224). Rather, percentages of agreement can be used to help identify the existence of widely varying patterns of rater judgments, both across WWYR subscales and across all rater pairs. No such widely varying patterns were found in the current study. The limitations of analyses involving percentages of agreement analysis were discussed by Abedi (1997), who argued that, although percentages of agreement can reveal the existence of widely varying patterns of agreement among raters, they can also yield different results from other analyses such as Pearson Product-Moment (PM) correlations.

Pearson Correlations

Pearson correlations were used to further examine Null Hypothesis 1. Table 4 contains the results of the Pearson correlations for WWYR rubric scoring across all rater pairs for the current study and for the Gearhart et al. study (1995). An examination of correlation scores for hypermedia narrative productions revealed that interrater reliability for four of the five WWYR subscales (Theme, Character, Setting and Plot) was comparable to the interrater reliability levels found in the Gearhart et al. (1995) WWYR reliability study for pen and paper-created narratives. For the fifth subscale (Communication), however, the correlational coefficient value was .16 higher in the Gearhart et al. study than in the current study. Despite the lower value found in the current study for Communication, Gearhart et al. related that an average subscale correlation higher than .50 could be considered adequate for a rubric such as the WWYR.

Correlations Between WWYR Subscales

Table 5 summarizes the comparison of WWYR correlations across all subscales for the current study and the Gearhart et al. (1995) study. The WWYR correlations observed for this study as well as the Gearhart et al. study demonstrated that ratings were highly correlated across all subscales. The r values were low for this study and for the Gearhart et al. (1995) and Novak et al. (1996) studies. However, set guidelines for what is an acceptable level of interrater reliability do not exist. Nonetheless, both Gearhart et. al and Novak et. al, whose studies analyzed holistic scores derived from the combined r values of Theme, Character, Plot, Setting and Communication, argued that r values which fell within the .50 to .70 range were acceptable for analytic writing rubrics.

In the current study the interrater reliability for Theme, Character, Plot and Communication subscales fell within the .50 and .59 range, but the level of interrater reliability (r=.49) for the Setting subscale did not. It is important to note that, in the Gearhart et al. study, a low coefficient value for the subscale of Setting was also found (r=.48).

A relatively small number of raters (n=5) was used in this study and the Gearhart et al. study, which may have contributed to the lower r values across all subscales. The attenuation of correlational coefficients may be another explanation for the low levels of interrater reliability (Gay, 1996). Accordingly, coefficients tend to be lower when a restricted range of values is utilized (e.g., the narrow range of only 3 out of a possible 6 WWYR subscale levels utilized by raters in this study). Thus, the more narrow the range of scores utilized by raters, the lower the coefficients. On the other hand, Gearhart et al. argued that if the number of raters was statistically were increased five-fold, r values in the .50 to .60 range for Theme, Character, Setting, Plot and Communication would be changed to .87, .89, .82, .86 and .89. Gearhart et al. used decision-study (multiplication of sample scores and aggregation of the results) coefficients to determine the number of raters needed to attain high reliability coefficients.

The acceptable interrater reliabilities for Theme, Character, Plot and Communication in this study were comparable to the acceptable levels found in the Gearhart et al. study, and the r values for the Setting subscale in both this study and the Gearhart et al. study were not acceptable. It is important to note that interrater reliability levels for Theme, Character and Plot in this study may have been lower (see Table 4.3) than the r values in the Gearhart et al. study because the researcher applied more stringent rating procedures in this study. Raters in the Gearhart et al. study were permitted to resolve differences greater than one scale point through discussion, whereas raters in this study were not permitted to resolve differences. In the current study all ratings were included in the final data set.

The r value for the Communication subscale in this study was considerably lower than the r value in the Gearhart et al. study (r=.50 versus .66). This sizable disparity, in the level of interrater reliability, may have been the result of the contrasting features of hypermedia created narrative productions versus pen and paper created narratives. The Communication subscale text primarily consists of evaluative prompts

designed to guide teachers in the assessment of writing style (See Table 2.1). Perhaps, in the current study raters' solely viewed textual features at the expense of the hypermedia features of graphics, sounds, buttons and scanned art.

Correlations Between WWYR Subscales

The highly correlated rater judgments, along all five WWYR subscales for the current study and for the Gearhart et al. study, provided further evidence of the reliability of WWYR raters' judgments. The true function of a writing rubric is that it "enables raters to apply standard criteria in making judgments about the quality of students' work" (Abedi, 1997, p. 8). Gearhart et al., Novak et al. and Abedi argued that highly correlated scores across rubric subscales can be viewed as a positive indication that raters' judgments are being consistently applied.

Results for Hotellings T^2 and One-Way MANOVA

Hotellings T^2

A Hotellings T^2 test was used to assess the differences across mean vectors of WWYR ratings for students in grades 2 and 3. No significant difference was observed between the mean vectors for grades 2 and 3, $F(1,54)=.87$, $p=.16$. Therefore, Null Hypothesis 2 was not rejected. Table 6 provides the descriptive statistics relative to Null Hypothesis 2. This test was also used to compare the ratings assigned to the productions of students in grade 2 with the ratings assigned to the productions of students in grade 3 as part of an effort to assess the developmental validity of WWYR scores for hypermedia created narrative productions. The developmental validity of scores generally corresponds to the positive correlation between students' chronological age and cognitive ability. Thus, it was assumed that older children were more capable of creating higher-level hypermedia narrative products than were younger children. Likewise, it was assumed that WWYR ratings in this study would reflect higher scores for grade 3 students than for students in grade 2. However, this assumption regarding the developmental validity of WWYR scores did not hold true for the creation of hypermedia narrative productions in the current study. This finding may be related to the narrow range of grade levels (grades 2 to 3) used in the current study. In contrast, Gearhart et al. (1995) and Novak et al. (1996) were able to successfully demonstrate the developmental validity of the WWYR for pen and paper samples due, in part, to the larger range of students (grades 1-6) who participated in their studies.

There was no significant difference observed across WWYR scores of students in grades 2 and 3. The insignificant difference may possibly have been due to classroom variables such as teacher-student interaction, computer availability and patterns of attendance. However, it should be noted that, observation of the two mean vectors revealed that grade 3 students scored higher than grade 2 students on all five subscales. Larger class sizes in the current study may have contributed to increased classroom noise levels, the teachers' inability to provide individualized instruction, and decreased time for individual students were allowed to use the computers. Also, patterns of average attendance over the length of the study indicated that students in grade 2 were present less often (n=20) than students in grade 3 (n=40).

The insignificant result obtained in the current study should function to guide the design of future studies that attempt to establish the validity of a measurement tool using developmental validity as a component. In order to establish developmental validity, a wide range of grade levels may be necessary to achieve significant results. Newman & Newman (1991), in their discussion on child development, pointed out that cognitive development does not necessarily have a perfectly linear relationship with chronological development. Thus, the argument for developmental validity can rest on a tenuous assumption if reasonably large grade/age level ranges are not used.

One-Way MANOVA

An examination of the results of the One-Way MANOVA conducted on the low, medium and high ability vectors of WWYR subscale scores indicated there was a statistically significant difference between the three ability groups ($F(2, 36)=2.59$, $p=.01$). Therefore, Null Hypothesis 3 was rejected. Table 7 provides descriptive statistics relative to Null Hypothesis 3. Table 8 provides an additional summary of these results across each of the five WWYR subscales.

Tukey HSD Tests were conducted on the mean vector scores of the three ability groups for all five WWYR subscales to follow-up these results. For the WWYR subscale of Theme, low-ability students (M=2.31, SD=.62) received lower scores than both medium-ability students (M=2.80, SD=.28) and high-ability students (M=2.86, SD=.31). For the WWYR subscales of Character, Setting, Plot and Communication, all differences were significant (i.e., low-ability students' scores were significantly lower than medium-ability students' scores, which were significantly lower than the high-ability students' scores).

The significant differences revealed between low, medium and high ITBS groups and WWYR subscale scores provided evidence for the sensitivity of the WWYR to the development of students' hypermedia/writing competence. The significant results of the One-Way MANOVA provided evidence that raters' judgments were evaluating students' skills as message-producers (communication through text and other meaning-based symbol systems). Dauite and Morse (1994), who used a similar curriculum (hypermedia/writing) in their study, found that students who were given the opportunity to compose in hypermedia were engaged in problem solving as they expressed themselves through the manipulation of a variety of meaning-based symbol systems, including text. Dauite and Morse drew the conclusion that students' hypermedia productions represented significant problem solving efforts, similar to what is required in process writing environments. The One-Way MANOVA did not yield results that would enable the researcher to directly describe the degree of relatedness of raters' WWYR judgments and students' ITBS scores. In order to describe the relationship between WWYR scores and literacy skill (as measured by the ITBS), an additional analysis was conducted.

Additional Findings

Correlations: WWYR Average Score and ITBS-NPR Score

The observed Pearson r correlation revealed a positive relationship between students' average WWYR score (averaged across the subscales of Theme, Character, Setting, Plot and Communication) and their ITBS National Percentile Rank (literacy skills score), $r = .83$, $p < .001$. The positive correlation ($r = .83$) between students' WWYR scores and ITBS scores revealed in this analysis provided evidence for the concurrent validity (the degree to which test scores are related to the scores on an already established test) of WWYR raters' judgments of hypermedia productions. According to Messick (1992) establishment of the concurrent validity of a measure can be a stepping-stone toward establishment of the content-related validity (the degree to which scores evaluate the specific domain they were designed to evaluate) of a measure. Hence, the developmental and concurrent validities established for WWYR raters' judgments of hypermedia productions represented an important initial attempt toward eventually establishing the content-related validity of the WWYR when applied to hypermedia productions.

The strongly positive linear relationship between ITBS literacy skill scores and WWYR rater judgments of hypermedia productions indicated that the hypermedia/writing curriculum used in the current

study involved literacy-based activities. The fact that students in this study expressed themselves through hypermedia features, and not solely text, indicated that students' literacy skill can be enhanced through student expression via hypermedia and multimedia features. (See Table 9 which provides additional information concerning students' utilization of hypermedia/writing features). This finding supports the claims of Daiute and Morse (1994), who observed that students' engaged in hypermedia writing developed literacy skill through the manipulation of text and other symbols. A weakness of the developmental and concurrent validity analyses was that evidence for obtaining the degree to which rater judgments of students' hypermedia productions evaluated textual features as well as textual and other hypermedia features (audio, hypermedia links, graphics, etc.) could not determined.

Educational Importance

The results of this study produced several important implications for the assessment of students' hypermedia products. According to Gearhart et. al (1995), reliable and valid assessment serves two general purposes: (a) to enhance classroom instruction (value), and (b) to inform educational policy (utility). The positive results yielded in this study concerning the reliability and validity of the WWYR suggest that teachers may benefit from applying WWYR assessment to their students' hypermedia narrative productions. The value of an assessment, is the degree to which it enhances teacher instruction by linking teachers' comments to their instructional objectives (Wolf & Gearhart, 1994). Accordingly, in order for teachers to properly evaluate both student outcomes and the instructional effectiveness of their hypermedia/writing curricula it would be advisable to use a reliable and valid instrument. Furthermore, the positive correlation between the students' ITBS literacy skill score and WWYR average score for hypermedia productions indicated that students who were engaged in a hypermedia/writing curriculum may have their literacy skills enhanced.

There are two implications for the large-scale application of WWYR results for hypermedia. First, the low reliabilities revealed in this study, although acceptable for classroom use, may not be appropriate for large-scale assessment. Additionally, the unacceptable reliability found for the Setting subscale matched the Gearhart et al. (1995) finding and provides further evidence that the WWYR, for both pen and paper and hypermedia, needs to be improved in order to be a reliable large-scale measure. Second, the content-related validity of the WWYR for hypermedia was not completely established. Messick (1992)

argued that several validity types must be evaluated in order for a measure to have content-related validity. Developmental and concurrent validity represented two lesser validity types in his hierarchy. In order for the WWYR to be used for large-scale assessment of students' hypermedia products, other types of validity should be documented as well.

References

Ayersman, D.J. (1996). Reviewing the research on hypermedia-based learning. Journal of Research on Computing in Education, 28(4), 500-526.

Gay, L.R. (1996). Educational Research. (5th ed.). New Jersey: Prentice-Hall.

Gearhart, M., Herman, J.L., Novak, J.R., & Wolf, S.A. (1995). Toward the instructional utility of large-scale writing assessment: Validation of a new narrative rubric. Assessing Writing, 2(2), 207-242.

Kinzer, C., & Leu, D.J. (1997). Focus on research: Exploring literacy and learning in electronic environments. Language Arts, 74, 126-135.

Linn, R., & Wilson, V. (1990). Review of the Iowa Test of Basic Skills form J. In J. Conoley and J. Impara (Eds.), Mental Measurement Yearbook (9th ed.). University of Nebraska Press.

McLellan, H. (1992, April). Narrative and episodic story structure in interactive stories. Paper presented at the annual meeting of the Association for Educational Communications and Technology, Washington, DC. (ERIC Document Retrieval Service No. ED 348 012).

Messick, S. (1992). Validity of test interpretation and use. In M. Alkin (Ed.), Encyclopedia of education research (6th ed., pp. 1487-1495). New York: Macmillan.

Newman, B., & Newman, P. (1991). The development through life perspective. Los Angeles, California: Brooks/Cole Publishing.

Novak, J.R., Herman, J.L., & Gearhart, M. (1996). Establishing validity for performanced-based assessments: An illustration for collections of student writing. Journal of Educational Research, 89(4), 220-233.

Novak, J.R., Herman, J.L., & Gearhart, M. (1996). Establishing validity for perfomanced-based assessments: An illustration for collections of student writing. Journal of Educational Research, 89(4), 220-233.

Palumbo, D.P., & Prater, D. (1993). The role of hypermedia in synthesis writing. Computers and Composition, 10(2), 59-70.

Reed, W.M. (1996). Assessing the impact of computer-based writing instruction. Journal of Research on Computing in Education, 28(4), 418-437.

Sharp, D., Kinzer, C., & Risko, V., and the Cognition and Technology Group at Vanderbilt. (1994, December). The young children's literacy project: Video and software tools for accelerating literacy in at-risk children. Presented at the annual meeting of the National Reading Conference, San Diego, CA.

Smith, K.J. (1992). Using multimedia with Navajo children: An effort to alleviate problems of cultural learning style, background of experience, and motivation. Reading & Writing Quarterly, 8, 287-294.

Swan, K., & Meskill, C. (1996). Using hypermedia in response-based literature classrooms: A critical review of commercial applications. Journal of Research on Computing Education, 29(2), 167-192.

The CRESST Line. (1997). Interrater reliability for performanced-based assessment. [Brochure]. Los Angeles, CA: Abedi, J.

Wagner, R. (1995). Hyperstudio Workshop 3.0. El Cajon, CA: Roger Wagner Publishing.

Wolf, S.A., & Gearhart, M. (1993a, 1993b). Writing What You Read: Assessment as a learning event (CSE Tech. Rep. No. 358). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Wolf, S.A., & Gearhart, M. (1994). Writing What You Read: A framework for narrative assessment. Language Arts, 7(6), 425-445.

Yang, S.C. (1996). A dynamic reading-linking-to-writing model for problem solving within a constructive hypermedia learning environment. Journal of Educational Multimedia and Hypermedia, 5(3/4), 263-283.

Table 1

Writing What You Read Narrative Analytic Rubric[2]

Theme	Character	Setting	Plot	Communication
Explicit-Implicit	Flat-Round	Backdrop-Essential	Simple-Complex	Context-bound
Didactic-Revealing	Static-Dynamic	Simple-Multi-functional	Static-Conflict	Literal-Symbolic
1: Not present or not developed through other narrative elements	1: One or two flat, static characters, with little relationship between characters	1: Backdrop setting with little or no indication of time or place ("There was a little girl. She like candy.")	1: One or two events with little or no conflict ("Once there was a cat. The cat liked milk.")	1: Writing bound to context ("You have to be there) and often dependent on drawing and talk to clarify the meaning
2: Meaning-centered in a series of list like statements ("I like my mom. And I like my dad. And I like my...")	2: Some rounding usually in physical description; relationship between characters is action driven	2: Skeletal indication of time and place often held in past time ("Once there was...") little relationship to other narrative elements	2: Beginning sequence of events, but out-of-sync occurrences; events without problem; problem without resolution	2: Beginning awareness of reader considerations; straightforward style and tone focused on getting the information out
3: Beginning statement of theme-often explicit and didactic ("The mean witch chased the children and she shouldn't have done that."	3: Continued rounding in physical description, particularly stereotypical features ("wart on the end of her nose")	3: Beginning relationship between setting and other narrative elements (futuristic setting to accommodate aliens and spaceships)	3: Single, linear episode with clear beginning, middle and end; the episode contains a problem, emotional response, action and outcome	3: Writer begins to make sense of explanations and transitions ("because" and "so"); literal style centers on description
4: Beginning revelation of theme on both explicit and implicit levels through the more subtle things characters say and do	4: Beginning insights into motivation and intention that drives the feeling and action of main characters often through limited omniscient point of view	4: Setting becomes more essential to the development of the story in explicit ways: characters may remark on the setting or the time and place may be integral to the plot	4: Plot increases in complexity with more than one episode; each episode contains problem, emotional response, action and outcome	4: Increased information and explanation for the reader (linking ideas as well as episodes); words more carefully selected to suit the narratives purpose
5: Beginning use of secondary themes, often tied to overarching theme, but sometimes tangential	5: Further rounding (in feeling and motivation); dynamic features appear in central characters and between characters	5: Setting may serve more than one function and the relationship between functions is more implicit and symbolic	5: Stronger relationships between episodes (with resolution in one leading to a problem in the next)	5: Some experimentation with symbolism (particularly figurative language) which shows reader considerations
6: Overarching theme multilayered and complex; secondary themes integrally related to the primary themes	6: Round, dynamic major characters through rich description of affect, intention and motivation	6: Setting fully integrated with the characters, action and theme	6: Overarching problem and resolution supported by multiple, episodes	6: Careful crafting of choices of story structure as well as vocabulary demonstrate considerate orchestration of all resources

BEST COPY AVAILABLE

[2] Some text from the WWYR has been removed from the original in order to fit the table on this page.

Table 2
Percentages of Agreement for all Five Subscales of the
WWYR Rubric Averaged Across Ten Pairs of Raters

WWYR Subscale	*±0*	*±1*	*n*
Theme	.70	.96	60
Character	.78	.99	60
Plot	.73	.99	60
Setting	.67	.99	60
Communication	.68	.99	60

Table 3

Percentages of Agreement for the WWYR Rubric Averaged Across All Subscales

WWYR Rating Material and Grade	*±0*	*±1*	*n*
Hypermedia Narratives: Grades 2-3 Mott, 1998	.71	.98	60
Pen and Paper Narratives: Grades 1-6 Gearhart et al. 1995	.46	.96	120
Collections of Pen and Paper Narratives: Grades 2-5 Novak et al. 1996	.25	.94	52

Table 4

Average Pearson Correlations for WWYR Rubric Scoring Across Ten Pairs of Raters

WWYR Rating Material and Grade		*Theme*	*Character*	*Setting*	*Plot*	*Comm*
Hypermedia Narratives: Grades 2-3 Mott, 1998 (n=60)	r	.59	.55	.49	.50	.50
	SD	.25	.31	.25	.29	.24
Pen and Paper Narratives Grades 1-6 Gearhart et al. 1995 (n=120)	r	.64	.59	.48	.57	.66
	SD	.10	.10	.12	.14	.10

Table 5

Comparison of WWYR Subscale Correlations: Pen and Paper Versus Hypermedia

Subscale	*Theme*	*Character*	*Setting*	*Plot*	*Communication*
Mott--Hypermedia			Samples (n=60)		
Theme	--	.86*	.79*	.79*	.73*
Character	--	--	.74*	.74*	.76*
Setting	--	--	--	.75*	.68*
Plot	--	--	--	--	.78*
Communication	--	--	--	--	--
Gearhart et al.--Pen and Paper			Samples (n=120)		
Theme	--	.83*	.81*	.83*	.86*
Character	--	--	.82*	.87*	.86*
Setting	--	--	--	.73*	.86*
Plot	--	--	--	--	.85*
Communication	--	--	--	--	--

Note: *p<.001.

Table 6

Descriptive Statistics: WWYR Subscales Across Grade Level

Statistics	*Dependent Variables*					
	n	Theme	Character	Setting	Plot	Comm.
Mean Vectors						
Grade Level						
2	20	2.25	1.90	2.14	2.19	2.18
3	40	2.62	2.21	2.30	2.47	2.46
Variance-Covariance Matrix						
Theme		.25	.28	.19	.21	.23
Character		--	.31	.20	.22	.21
Setting		--	--	.25	.22	.17
Plot		--	--	--	.29	.23
Comm.		--	--	--	--	.24

Table 7
Descriptive Statistics: WWYR Subscales Across ITBS Ability Level

Statistics	*Dependent Variables*					
	n	Theme	Character	Setting	Plot	Comm.
Mean Vectors						
ITBS Ability Level						
Low	13	2.31	1.80	1.96	2.10	2.14
Medium	13	2.80	2.34	2.32	2.66	2.52
High	14	2.86	2.60	2.66	2.74	2.77
Variance-Covariance Matrix						
Theme		.21	.15	.19	.13	.11
Character		--	.23	.13	.15	.13
Setting		--	--	.19	.14	.11
Plot		--	--	--	.19	.14
Comm.		--	--	--	--	.16

Table 8

Mean WWYR Subscale Scores for Low, Medium and High

Ability Grade 3 Students

WWYR Subscale	*ITBS NPR/Literacy Category*	*Mean Score*	*SD*	*n*	*F*	*Sig*
Theme	Low	2.31	.62	16		
	Medium	2.80	.28	10	6.19	.01
	High	2.86	.31	13		
Character	Low	1.80	.50	16		
	Medium	2.34	.38	10	10.77	.01
	High	2.60	.51	13		
Setting	Low	1.96	.42	16		
	Medium	2.32	.56	10	9.34	.01
	High	2.66	.34	13		
Plot	Low	2.10	.54	16		
	Medium	2.66	.34	10	9.28	.01
	High	2.74	.34	13		
Communication	Low	2.14	.47	16		
	Medium	2.52	.34	10	9.20	.01
	High	2.77	.35	13		

Table 9-Frequency of Hyperstudio Multimedia Features
Used in Students' Hypermedia Narrative Productions

	Hypermedia/Multimedia Feature						
Grade Level	Button with Hypermedia Link	Button with Audio	Button with Video	Text Box	Graphics Text	Scanned Art	Graphics Objects (Clip Art)
2 (n=20)	100%	81%	0%	100%	45%	96%	82%
3 (n=40)	100%	100%	5%	100%	64%	100%	100%

Note. In three out of the four classrooms where hypermedia/writing occurred students' use of hypermedia/multimedia features was controlled by the teachers.

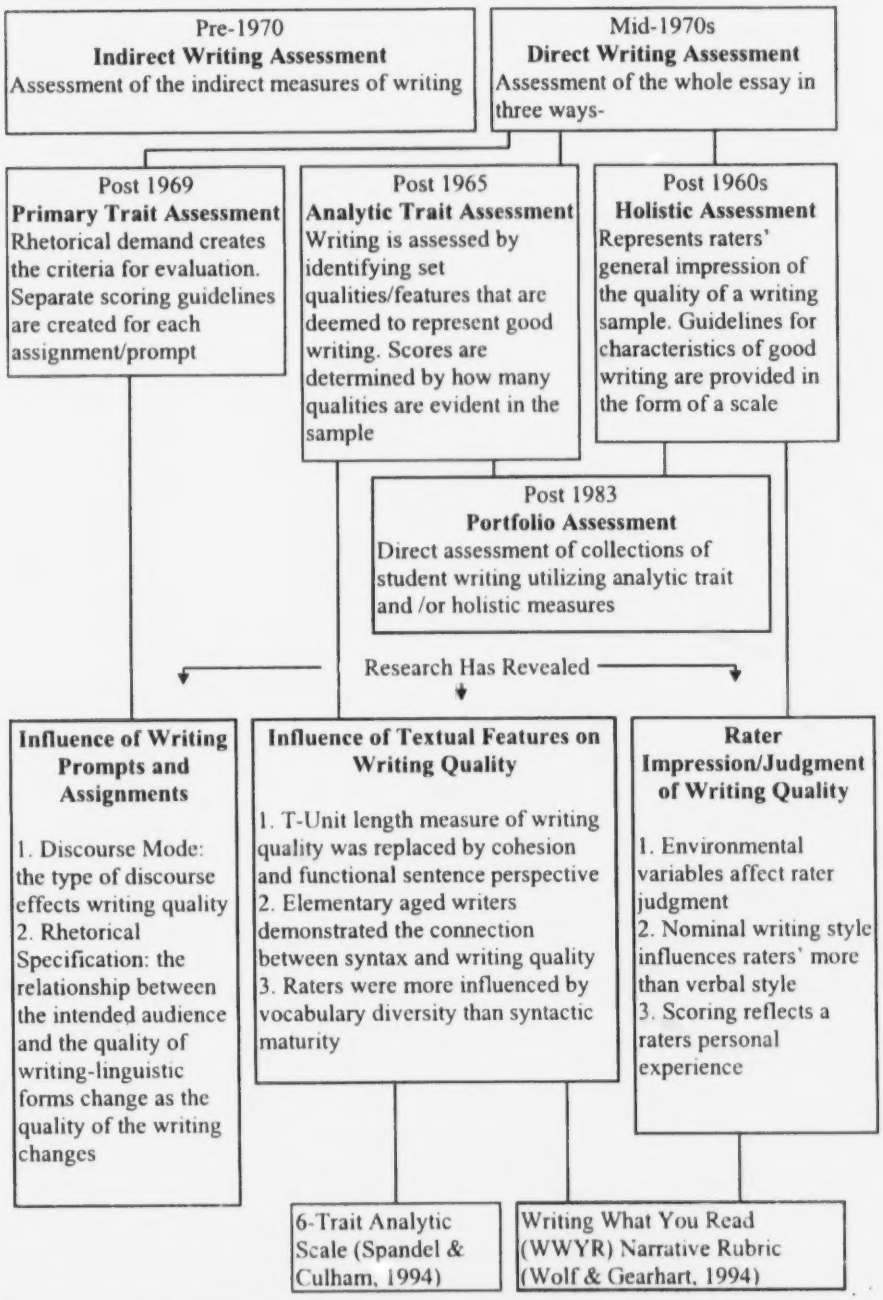
Pre-1970
Indirect Writing Assessment
Assessment of the indirect measures of writing
Mid-1970s
Direct Writing Assessment
Assessment of the whole essay in three ways-
Post 1969
Primary Trait Assessment
Rhetorical demand creates the criteria for evaluation. Separate scoring guidelines are created for each assignment/prompt
Post 1965
Analytic Trait Assessment
Writing is assessed by identifying set qualities/features that are deemed to represent good writing. Scores are determined by how many qualities are evident in the sample
Post 1960s
Holistic Assessment
Represents raters' general impression of the quality of a writing sample. Guidelines for characteristics of good writing are provided in the form of a scale
Post 1983
Portfolio Assessment
Direct assessment of collections of student writing utilizing analytic trait and /or holistic measures
Research Has Revealed
Influence of Writing Prompts and Assignments
1. Discourse Mode: the type of discourse effects writing quality
2. Rhetorical Specification: the relationship between the intended audience and the quality of writing-linguistic forms change as the quality of the writing changes
Influence of Textual Features on Writing Quality
1. T-Unit length measure of writing quality was replaced by cohesion and functional sentence perspective
2. Elementary aged writers demonstrated the connection between syntax and writing quality
3. Raters were more influenced by vocabulary diversity than syntactic maturity
Rater Impression/Judgment of Writing Quality
1. Environmental variables affect rater judgment
2. Nominal writing style influences raters' more than verbal style
3. Scoring reflects a raters personal experience
6-Trait Analytic Scale (Spandel & Culham, 1994)
Writing What You Read (WWYR) Narrative Rubric (Wolf & Gearhart, 1994)

Figure 1: Writing Assessment Research: Definitions, Time Table and Sample Instruments

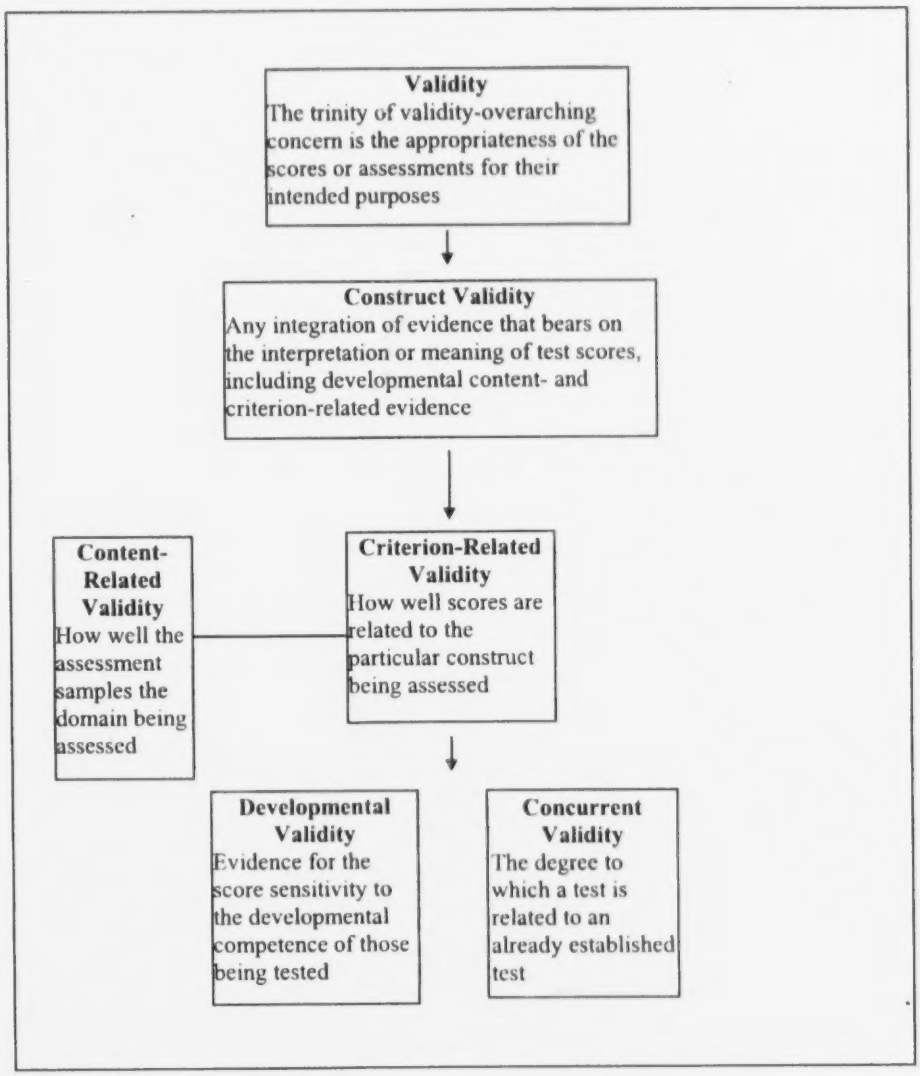
Validity
The trinity of validity-overarching concern is the appropriateness of the scores or assessments for their intended purposes
Construct Validity
Any integration of evidence that bears on the interpretation or meaning of test scores, including developmental content- and criterion-related evidence
Content-Related Validity
How well the assessment samples the domain being assessed
Criterion-Related Validity
How well scores are related to the particular construct being assessed
Developmental Validity
Evidence for the score sensitivity to the developmental competence of those being tested
Concurrent Validity
The degree to which a test is related to an already established test

Figure 2: Messick's Hierarchical and Expanded Definition of Validity

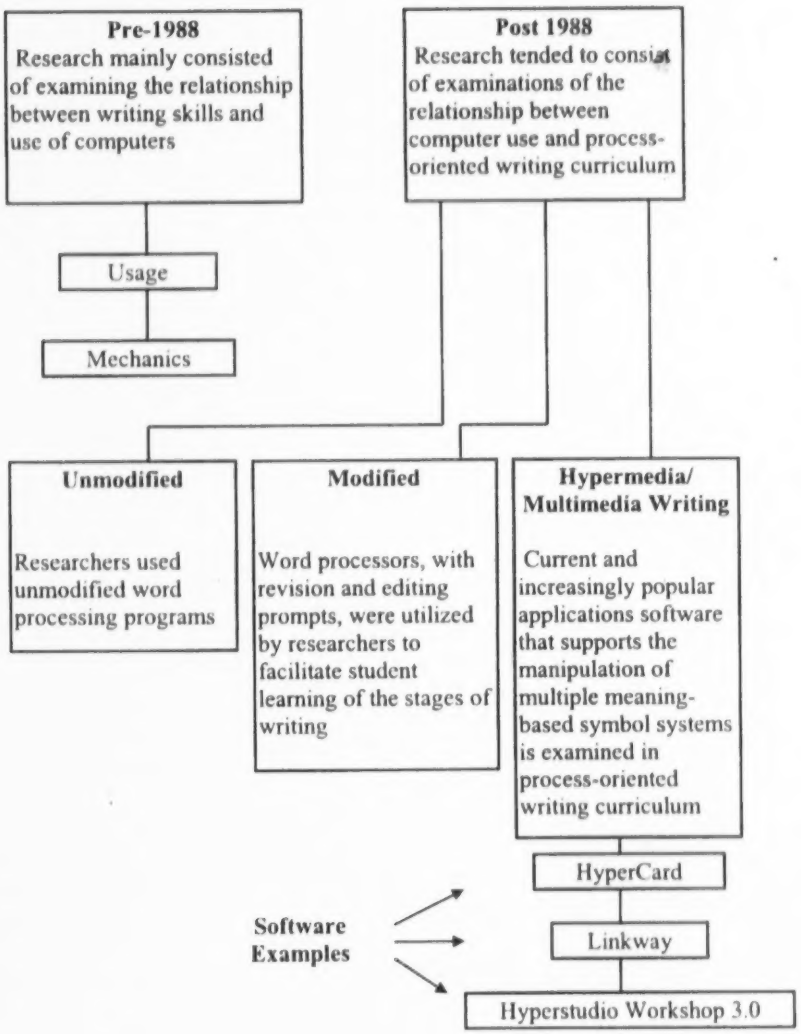
Pre-1988
Research mainly consisted of examining the relationship between writing skills and use of computers
Usage
Mechanics
Post 1988
Research tended to consist of examinations of the relationship between computer use and process-oriented writing curriculum
Unmodified
Researchers used unmodified word processing programs
Modified
Word processors, with revision and editing prompts, were utilized by researchers to facilitate student learning of the stages of writing
Hypermedia/ Multimedia Writing
Current and increasingly popular applications software that supports the manipulation of multiple meaning-based symbol systems is examined in process-oriented writing curriculum
Software Examples
HyperCard
Linkway
Hyperstudio Workshop 3.0

Figure 3: Research on Computer-based Writing Instruction